DOCUMENT RESUME

ED 389 753 TM 024 386

AUTHOR Veerkamp, Wim J. J.; Berger, Martijn P. F.
TITLE Some New Item Selection Criteria for Adaptive Testing. Research Report 94-6.
INSTITUTION Twente Univ., Enschede (Netherlands). Faculty of Educational Science and Technology.
PUB DATE Nov 94
NOTE 38p.
AVAILABLE FROM Bibliotheek, Faculty of Educational Science and Technology, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Ability; *Adaptive Testing; *Bayesian Statistics; Computer Assisted Testing; *Criteria; *Estimation (Mathematics); Foreign Countries; Item Response Theory; Maximum Likelihood Statistics; *Selection; Simulation; *Test Items
IDENTIFIERS Weighting (Statistical)

ABSTRACT

In this study some alternative item selection criteria for adaptive testing are proposed. These criteria take into account the uncertainty of the ability estimates. A general weighted information criterion is suggested of which the usual maximum information criterion and the suggested alternative criteria are special cases. A simulation study was conducted to compare the different criteria. The results showed that the likelihood weighted mean information criterion was a good alternative to the maximum information criterion. Another good alternative was a maximum information criterion with the maximum likelihood estimate of ability replaced by the Bayesian EAP estimate. An appendix discusses the interval information criterion for the two- and three-parameter logistic item response theory model. (Contains 5 figures and 15 references.) (Author)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 389 753

Some New Item Selection Criteria for Adaptive Testing

Research Report 94-6

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☒ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
J. NELISSEN
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Wim J.J. Veerkamp

Martijn P.F. Berger

BEST COPY AVAILABLE

faculty of
EDUCATIONAL SCIENCE AND TECHNOLOGY

University of Twente

Department of
Educational Measurement and Data Analysis

TM024386

Some New Item Selection Criteria for Adaptive Testing

Wim J.J. Veerkamp
Martijn P.F. Berger

Some new item selection criteria for adaptive testing, Wim J.J. Veerkamp and Martijn P.F. Berger - Enschede: University of Twente, Faculty of Educational Science and Technology, November, 1994, - 32 pages.

Abstract

In this study some alternative item selection criteria for adaptive testing are proposed. These criteria take into account the uncertainty of the ability estimates. A general weighted information criterion is suggested of which the usual maximum information criterion and the suggested alternative criteria are special cases.

A simulation study was conducted to compare the different criteria. The results showed that the likelihood weighted mean information criterion was a good alternative to the maximum information criterion. Another good alternative was a maximum information criterion with the maximum likelihood estimate of ability replaced by the Bayesian EAP estimate.

Key words: item selection, adaptive testing, test design, efficiency

Some new item selection criteria for adaptive testing

Introduction

Adaptive testing was proposed and developed by Lord (1971, 1980) and Weiss (1976, 1978), among others, to overcome the disadvantages of standardized tests. In a standardized test the different abilities cannot always be estimated with equal precision. The objective of both adaptive and standardized testing is to estimate the ability of examinees as efficiently as possible with a minimum number of items. The main difference between a standard test and an adaptive test is that in a standard test items are selected beforehand, while in an adaptive test items are selected sequentially during the administration of the test. Essentially, this latter procedure will lead to testing sessions where different examinees take different forms of a test.

In Item Response Theory (IRT), a main concept used for the selection of items for inclusion in a test is Fisher's information function. Usually, it is assumed that a large number of items has been calibrated and collected in an item bank. From the model describing the characteristics of these items, it is possible to derive how much information each item will produce. It can be shown (Bradley & Gart, 1962) that for known item parameters the Maximum Likelihood Estimator $\hat{\theta}_{ML}$ of the ability θ is consistent and asymptotically normally distributed with variance equal to $I(\theta)^{-1}$, i.e.

$$\hat{\theta}_{ML} \sim AN(\theta, I(\theta)^{-1}), \tag{1}$$

where $I(\theta)$ is Fisher's information, which is defined as:

$$I(\theta) \equiv E\left[\left(\frac{d \ln(L_n(\theta; \boldsymbol{x}_n))}{d\theta}\right)^2\right], \tag{2}$$

with $L_n(\theta; \boldsymbol{x}_n)$ being the likelihood function belonging to the observations $\boldsymbol{x}_n = [x_1,....x_n]$, where n is the number of items the examinee has already answered, and x_j is the score on item j: $x_j = 0$ if item j is answered wrong and $x_j = 1$ if it is answered correct. $L_n(\theta; \boldsymbol{x}_n)$ is given by:

$$L_n(\theta; \boldsymbol{x}_n) = \prod_{j=1}^{n} L(\theta; x_j) = \prod_{j=1}^{n} P_j(\theta)^{x_j}[1 - P_j(\theta)]^{1-x_j}. \tag{3}$$

In equation (3) $P_j(\theta)$ denotes the probability that an examinee with ability θ answers item j correct. An example of this item characteristic function is given in equation (23) of the appendix. In IRT models $I(\theta)$ is usually referred to as the test information function. An important feature of the test information function is that it consists entirely of independent and additive contributions from each of n separate items (Lord, 1980, p.72), i.e.

$$I(\theta) = \sum_{i=1}^{n} I_i(\theta), \tag{4}$$

where

$$I_i(\theta) = \frac{(dP_i(\theta)/d\theta)^2}{P_i(\theta)[1-P_i(\theta)]} \qquad (5)$$

is the item information function for dichotomous IRT models. It should be emphasized, that the use of Fisher's information is only valid asymptotically; in small samples the inverse of the information function is only a lower bound for the actual variance of the ability estimator.

Several measures for the selection of items in adaptive testing have been proposed; Kingsbury & Zara (1989) give a review of them. A widely used criterion is Lord's (1977) maximum information criterion, which uses the maximum likelihood (ML) estimate of the ability $\hat{\theta}_{ML}$. There are, however, two problems connected with the ML estimation of ability during an adaptive test.

The first problem is that the likelihood function does not have a finite maximum in case of a zero or perfect score. In practice, this omission may lead to using an arbitrary extreme value on the ability scale as an estimate when all responses are correct or all are wrong.

A second problem was identified by Samejima (1973). She located multiple solutions to the likelihood equations (multiple maxima in the likelihood function) when an examinee's ability is estimated with the three-parameter logistic model (see also Yen, Burket & Sykes, 1991). Although in the practical applications studied by Lord (1980, p.59), multiple solutions did not occur when the number of items was larger than 20, this problem of multiple local maxima will often arise in adaptive testing where the number of selected items tends to be

small.

One solution to the above mentioned problems with ML estimation is to replace it by EAP estimation with a uniform prior. This method, always gives a finite and unique ability estimate. The EAP estimator with uniform prior is defined as:

$$\hat{\theta}_{EAP} \equiv \int_{-\infty}^{\infty} \theta L_n(\theta; x_n) d\theta \Big/ \int_{-\infty}^{\infty} L_n(\theta; x_n) d\theta . \qquad (6)$$

In practice, however, this estimator can only be approximated by numerical integration (see equation (21)). See Bock & Mislevy (1982) for more details on the EAP-estimator.

Another solution to the problem with ML estimation is to use an item selection criterion which does not need ability estimates in each step of the adaptive testing process. One of the new criteria proposed in this paper possesses this feature.

In this paper a general item selection criterion will be proposed. It will be shown that the usual maximum information criterion and the new item selection criteria proposed in this paper are special cases of this general criterion. These criteria will be compared with each other. Since an analytical comparison is not feasible, the performances of these criteria were investigated by means of simulated data. In the following section the proposed measures for item selection will first be described. Their advantages and disadvantages will then be discussed. Finally, the results of a simulation study will be presented, and some concluding remarks about the relative merits of each measure will be made.

Criteria for item selection

In this section a number of criteria for the selection of items will be described. In adaptive testing the ability of an examinee is (re)estimated during the administration of the test. Such provisional estimates are then used to select each next item. For one of the following criteria, however, item selection is done without the need to use such provisional estimates.

First, however, a general formulation of an item selection criterion will be presented. The item selection criteria discussed in this paper can all be seen as special cases of this general criterion, which will be called the General Weighted Information Criterion.

General Weighted Information Criterion

In most current adaptive testing procedures the item with the highest value of the information function for the ability parameter estimate is usually selected. The highest information for the ability parameter estimate, however, may not be the highest for the true parameter value, and the difference between the ability estimate and the true value may disturb the results. Instead of just using a single value of the information function, the information function values for the whole range of values on the ability scale can be used. This range can be transformed into a single value by a weighted average. This general idea of selecting items by means of the largest weighted mean information over the whole range of abilities can be formulated as:

$$\max_{i \in \mathbf{I}_n} \int_{-\infty}^{\infty} W_n(x_n;\theta)\, I_i(\theta)\, d\theta\,, \tag{7}$$

or for a discrete ability scale:

$$\max_{i \in \mathbf{I}_n} \sum_{j=1}^{k} W_n(x_n;\theta_j)\, I_i(\theta_j), \tag{8}$$

where $\mathbf{I}_n$ denotes the total set of items from which the items are to be selected, i.e. the item pool minus the n items that have already been selected. Whether formulation (7) or (8) is used depends on the weight function $W_n(x_n;\theta)$. For probability mass functions, for example, only formulation (8) can be used. When formulation (8) is used as a numerical approximation of formulation (7) θ_j is one of k quadrature points in a numerical integration procedure.

The general formulation in (7) and (8) includes a variety of criteria. Which specific criterion is used depends on the weight function. The weight function can be used in at least three different ways. Firstly, when a test is designed to select persons for placement or admission, high information is often needed for some prespecified values on the ability scale. A weight function with peaks at these values may then be appropriate. Secondly, auxiliary information about an examinee can be implemented by using a prior distribution of abilities $g(\theta)$ as a weight function, i.e. $W_n(x_n;\theta) = g(\theta)$. An application of such weights is given by Berger (to appear). Finally, the uncertainty of the ability estimate can be taken into account by using special weight functions. In this paper the latter way of using weights is stressed.

Point Information Criterion

The most commonly used heuristic measure for the selection of items is the so-called maximum information criterion (Lord, 1977). We will refer to this criterion as the *Point Information Criterion* with ML estimation:

$$\max_{i \in \mathbf{I}_n} I_i(\hat{\theta}_{ML}). \qquad (9)$$

The application of this criterion assumes a provisional estimate of θ. The following sequential estimation procedure is usually applied. The procedure may start with an initial estimate $\hat{\theta} = 0$. The next estimate will then be a positive number for a correct and a negative number for an incorrect answer to the first item, and the most appropriate point estimate will then be located at the extreme ends of the ability scale, because the maximum of the likelihood function for a perfect or zero score is plus or minus infinity, respectively. As soon as the total score of an examinee is not perfect or zero anymore, regular ML estimation can be applied. As mentioned earlier, this estimation procedure does not always proceed without problems.

The Point Information Criterion can also be based with other estimators, such as the EAP estimator $\hat{\theta}_{EAP}$. The EAP estimation procedure can be used for the whole test without problems, because it always gives finite and unique estimates.

Note, that this criterion fits into the general formulation of equation (8); the weight function reduces to 1 for $\theta = \hat{\theta}$ and 0 otherwise:

$$W_n(x_n;\theta) = J_{\{\hat{\theta}\}}(\theta), \qquad (10)$$

where $J_{[.]}(.)$ denotes an appropriate indicator function (see e.g. Mood, Graybill and Boes, 1974, p. 20), and $\hat{\theta}$ may denote any estimator.

Interval Information Criterion

An objection against using the Point Information Criterion, is that it does not take into account the uncertainty in the estimates in each step. Maximum information at an ability estimate is not always maximum at the true ability value. To overcome this problem, a criterion based on the value of the information function for a certain interval of θ-values is proposed. This so-called *Interval Information Criterion* selects the item with the highest mean value of the information function in a confidence interval of θ:$[\hat{\theta}_L, \hat{\theta}_R]$. The size of the confidence interval expresses the uncertainty of the estimator. Since the estimator $\hat{\theta}_{ML}$ is asymptotically normally distributed with mean θ and variance $I(\theta)^{-1}$, a corresponding confidence interval can be formulated as:

$$\left[\hat{\theta}_L = \hat{\theta}_{ML} - \lambda\, I(\hat{\theta}_{ML})^{-1/2} \,,\; \hat{\theta}_R = \hat{\theta}_{ML} + \lambda\, I(\hat{\theta}_{ML})^{-1/2} \right],$$

where $\lambda = \Phi^{-1}[(1+\gamma)/2]$ determines the size of the $100\gamma\%$-confidence interval, and $\Phi(.)$ is the standard normal cumulative distribution function, and γ the confidence coefficient.

For the EAP estimator a $100\gamma\%$-confidence interval can be formulated as:

$$[\hat{\theta}_L = \hat{\theta}_{EAP} - \lambda\sqrt{Var(\theta \mid x)}\,, \hat{\theta}_R = \hat{\theta}_{EAP} + \lambda\sqrt{Var(\theta \mid x)}\,], \qquad (12)$$

where $Var(\theta \mid x)$ is the Bayesian posterior variance with a uniform prior:

$$Var(\theta \mid x) = \frac{\int_{-\infty}^{\infty} (\theta - \hat{\theta}_{EAP})^2 L_n(\theta; x_n) d\theta}{\int_{-\infty}^{\infty} (L_n(\theta; x_n) d\theta} . \qquad (13)$$

Here it is assumed that the EAP estimator is approximately normally distributed with mean θ and variance $Var(\theta \mid x)$.

The Interval Information Criterion can now be formulated as the area under the information function from the left boundary $\hat{\theta}_L$ to the right boundary $\hat{\theta}_R$ of the interval, i.e.:

$$\max_{i \in \mathbf{I}_n} \int_{\theta=\hat{\theta}_L}^{\hat{\theta}_R} I_i(\theta) d\theta . \qquad (14)$$

It should be noted that a mean value can be obtained by dividing the formula by the length of the confidence interval. However, this length is the same for all items. Hence, for the selection of items this constant term can be left out. Note also, that this equation can be obtained from the general formulation (7) by using a weight function which is 1 for parameter values within the confidence interval and 0 outside that interval:

$$W_n(x_n;\theta) = J_{[\hat{\theta}_L,\hat{\theta}_R]}(\theta), \tag{15}$$

with $J_{[..]}(.)$ again denoting an indicator function.

This criterion also needs a starting value, because a confidence interval cannot be constructed when an estimate of θ is not available, or when the total score is zero or perfect. This problem can be solved by starting with an unbounded interval:

$$\max_{i \in \mathbf{I}_n} \int_{\theta=-\infty}^{\infty} I_i(\theta)d\theta . \tag{16}$$

So, in fact $\hat{\theta}_L = -\infty$ and $\hat{\theta}_R = \infty$ are now used as starting values. These starting values may be used in the selection process as long as the total score is zero or perfect.

Likelihood Weighted Information Criterion

Another criterion proposed in this study will be called the *Likelihood Weighted Information Criterion*, because the likelihood function is used as a weight function, i.e.:

$$W_n(x_n;\theta) = L_n(\theta;x_n) . \tag{17}$$

This choice amounts to giving more weight to the information function value $I(\theta)$ when it is more likely that the corresponding θ-value on the θ-scale is the true ability value of the examinee.

The Likelihood Weighted Information Criterion is defined as the area under a function that is a product of the likelihood function and the information function:

$$\max_{i \in \mathbf{I}_n} \int_{-\infty}^{\infty} L_n(\theta; \boldsymbol{x}_n) I_i(\theta) d\theta . \qquad (18)$$

It should be noted, that a weighted mean can be obtained by dividing this formula by $\int_{-\infty}^{\infty} L_n(\theta; \boldsymbol{x}_n) d\theta$, which is again constant for all items and can be left out in the selection process. The same starting value as the one proposed for the Interval Information Criterion can be applied in this case, but with a different argument, namely $L_0(\theta; \boldsymbol{x}_0) = 1$, because all values on the θ-scale can be assumed to be equally likely when no data are available and no prior distribution is formulated.

A major advantage of the present criterion is that provisional ability estimates are not needed anymore. Ability is estimated only once, namely at the end of the test administration. Hence, the problems of ML estimation of ability during the test are circumvented. Furthermore, the Likelihood Weighted Information Criterion shares the advantage with the Interval Information Criterion that the first item can be chosen on the basis of an average information. This choice is more appealing than starting with the information of an arbitrary ability value $\hat{\theta}=0$.

Summary remarks

Our conjecture is that the Likelihood Weighted Information Criterion and the Interval Information Criterion will perform better than the Point Information Criterion, because they take the uncertainty of the ability estimate into account. Since the likelihood function reveals how much weight should be attached to the information function value at all possible values on the ability scale, it can be inferred that the Likelihood Weighted Information Criterion will perform better than the Interval Information Criterion, because the weight function connected with the Interval Information Criterion is constant for the whole interval.

To show how the weight functions of these three item selection criteria differ they are presented in Figure 1. The weight functions are taken from a test with a length of $n = 10$ items administered to an examinee with ability equal to $\theta = 3$. The ability estimate in this figure is $\hat{\theta} = 2.6$ and the confidence interval is $[\hat{\theta}_L, \hat{\theta}_R] = [1.66, 3.54]$.

Insert Figure 1 about here

A Simulation Study

Simulated data, based upon the three parameter logistic IRT model, were used to compare the performances of the different criteria. An adaptive testing process was simulated for the following five criteria; two versions of the Point Information Criterion, one with ML estimation and the other with EAP estimation

of the ability; two versions of the Interval Information Criterion, also both with ML and EAP; and the Likelihood Weighted Information Criterion.

Method

Since the computation of the proposed criteria is rather complicated, some approximations were used. These approximations used the set of θ-values ranging from -5.0 to 5.0 with steps of 0.1, i.e. the set defined as $\Theta = \{\theta : \theta = -5.0 + k/10 \ \forall \ k \in \mathbf{N}, \ k \leq 100\}$, where $\mathbf{N}$ denotes the set of natural numbers.

To avoid any problems with multiple local maxima of the likelihood function, the global maximum of the likelihood function was found each time by a search through all the θ-values from Θ; that is, our ML estimator was defined as:

$$\hat{\theta}_{ML} \in \Theta: \ L_n(\hat{\theta}_{ML}; \boldsymbol{x}_n) \geq L_n(\theta; \boldsymbol{x}_n) \quad \forall \ \theta \in \Theta. \tag{19}$$

Since the integral used in the Interval Information Criterion in (14) has a closed form solution for the three-parameter logistic model (see equation (25) in the appendix), no approximation of this criterion is needed. On the other hand, the integral in the equation for the Likelihood Weighted Information Criterion does not have a closed form solution, except for the first item. To solve this problem, the function was integrated numerically with quadrature points in Θ. In particular, the function was approximated by a step function within the interval [-5.0, 5.0] with steps of 0.1. These steps were sufficiently small to obtain good approximations. Such step functions can easily be integrated by summation (see e.g. Apostol, 1967, pp. 64-69). So, the Likelihood Weighted Information Criteri-

on (18) was approximated by:

$$\max_{i \in \mathbf{I}_n} \sum_{\theta \in \Theta} L_n(\theta; \boldsymbol{x}_n) I_i(\theta). \tag{20}$$

The integrals of the EAP-estimator given in equation (6), and in the posterior variance in equation (13) were also replaced by sums, i.e.:

$$\hat{\theta}_{EAP} = \sum_{\theta \in \Theta} \theta L_n(\theta; \boldsymbol{x}_n) \Big/ \sum_{\theta \in \Theta} L_n(\theta; \boldsymbol{x}_n), \tag{21}$$

and

$$Var(\theta \mid x) = \frac{\sum_{\theta \in \Theta} (\theta - \hat{\theta}_{EAP})^2 L_n(\theta; x_n)}{\sum_{\theta \in \Theta} L_n(\theta; x_n)}. \tag{22}$$

Data Generation

Since the characteristics of the items in the item bank were expected to influence the performance of the criteria, items with a wide range of parameter values were simulated. The item parameters were drawn from uniform distributions, i.e. $a_i \sim U(0.5, 2.0)$, $b_i \sim U(-3.0, 3.0)$, and $c_i \sim U(0.15, 0.30)$. The scores x_j on item j were obtained by $x_j = J_{[0, P_j(\theta)]}(u)$, where u is randomly drawn from a $U(0, 1)$-distribution, and $P_j(\theta)$ is the three-parameter logistic function given by equation (23) in the appendix.

Simulated Conditions

The total number of items in the item bank may influence the results. Therefore, two different item bank sizes were assumed, namely an item bank with 200 items and one with 400 items. To investigate the effect of the abilities of the examinees, seven different θ-values were selected in the simulation, namely: -3, -2, -1, 0, 1, 2, and 3. Thus 14 different conditions were simulated in this study. The number of replications for each condition was R = 200. All tests had the same final test length, namely 60 items. Since some criteria were expected to give better results in the beginning of the test, the results of the criteria after the selection of 5, 10, ... , 55, and 60 selected items, respectively, are displayed.

Measures of Comparison

Different measures can be used to compare the performance of the item selection criteria. Measures used in this study were: (a) mean average item information: $\frac{1}{R}\sum_{r=1}^{R}\frac{I(\theta)_r}{n}$, where n is the number of selected items and θ is the true ability value; (b) minimum test information: $\min_r I(\theta)_r$, with the minimum taken over all R replications; (c) variance: $\frac{1}{R-1}\sum_{r=1}^{R}(\bar{\theta}-\hat{\theta}_r)^2$, where $\bar{\theta}=\frac{1}{R}\sum_{r=1}^{R}\hat{\theta}_r$ is the mean of the estimates $\hat{\theta}_r$ over R replications; (d) bias: $\theta-\bar{\theta}$; (e) mean squared error: $\frac{1}{R}\sum_{r=1}^{R}(\theta-\hat{\theta}_r)^2$; and (f) maximum squared error: $\max_r(\theta-\hat{\theta}_r)^2$. The estimates $\hat{\theta}_r$ were both ML and EAP estimates.

Results

Among the measures used to compare the performances of the different item selection criteria, the mean average item information and the minimum test information gave the most striking results. For each of the simulated conditions these results are presented in Figures 2 through 5. The numbers in the figures represent the five selection criteria. When these numbers are displayed close behind each other, the order in which they appear represents the order of the lines at that point. When the numbers in a graph are omitted, the lines are very close, which means that the differences were very small. The lines below all other lines in each graph represent the performance with random selection.

In Figures 2 and 3 the mean values of the average item information over the R = 200 replications for the criteria are presented, for item banks with 200 items and 400 items, respectively. The two Interval Information Criteria performed slightly worse than the two Point Information criteria and the Likelihood Weighted Information Criterion. The Likelihood Weighted Information Criterion and the Point Information Criterion with EAP estimation performed better than the Point Information Criterion with ML estimation. The differences were the most striking for extreme ability values. Moreover, the differences decreased with increasing test length. The typical form of most lines in Figures 2 and 3 can be explained as follows. In the beginning of a test item selection improves because more information becomes available as items are selected. But, after a while the item bank gets exhausted, because the best items have already been selected, and less informative items are left over to choose among.

Insert Figures 2 and 3 about here

In Figures 4 and 5 the smallest values of the test information function over the 200 replications are given, for the item banks with 200 and 400 items, respectively. It can be seen that the Interval Information Criteria and the Point Information Criterion with ML estimation often lead to lower informative tests than the Likelihood Weighted Information Criterion and the Point Information Criterion with EAP.

Insert Figures 4 and 5 about here

In summary, these results show that the best criteria seem to be the Likelihood Weighted Information Criterion and the Point Information Criterion with EAP estimation.

Discussion and Summary

In this paper some new item selection criteria for adaptive tests are proposed. These criteria are formulated as special cases of a more general item selection criterion, i.e. the General Weighted Information Criterion. This General Weighted Information Criterion not only includes the well-known maximum (Point) Information Criterion, but also includes two new criteria, namely the Interval Information Criterion and the Likelihood Weighted Information Criterion. Of course, other weight functions in the general formulation can also be considered.

The often used Point Information Criterion has some disadvantages. Firstly, the uncertainty of the ability estimate is not taken into account. Secondly, the ability has to be estimated during the test administration. This may become troublesome when maximum likelihood estimation is used. For a perfect or zero score the likelihood function has no finite maximum, whereas for the three-parameter logistic IRT model the likelihood equation may have more than one solution. Such problems, however, will not arise when EAP estimation is used.

The Interval Information Criterion takes the uncertainty of the ability estimate into account by using a weight function which is 1 for ability values within a confidence interval for the ability and 0, otherwise.

Another way to take the uncertainty of the ability estimate into account is to use the likelihood function as a weight function. This is done in the Likelihood Weighted Information Criterion. The Likelihood Weighted Information Criterion does not need ability estimates during the item selection process, and ability has to be estimated only once, namely at the end of the test administration. A disadvantage of this criterion, however, is that it needs more computer time than the other criteria. Our algorithm for the Likelihood Weighted Information Criterion took about 9 seconds CPU-time to select one item out of the 200-items bank, whereas the computation of the Point Information Criterion and the Interval Information Criterion needed about 1/10th of a second. A personal computer with an Intel 80386 processor and mathematical coprocessor, with 25 MHz clock speed and 4 MB internal memory, was used. The CPU-time was reduced by a factor 5 on a 486-50 MHz-PC. In practice, however, an examinee does not have to wait for such a long time for the next item. While the examinee works on the response to an item, the computer can do the calculations needed for the selection of the next item. Furthermore, faster computers will be available in the near

future, and the used algorithms can be improved and made more efficient.

To compare the selection procedures, a small simulation study was performed. In this study, the performances of two of the alternatives were somewhat better than the familiar Point Information Criterion with ML estimation. One of these alternatives uses an alternative estimator, namely the EAP estimator with a uniform prior. The other criterion with a promising performance is the Likelihood Weighted Information Criterion. The good performance of the Likelihood Weighted Information Criterion was as expected, and the good results for the Point Information Criterion with EAP can be explained by the global resemblance of the EAP estimator with the Likelihood Weighted Information Criterion. The EAP estimator with uniform prior can be seen as a likelihood weighted mean ability.

Appendix

The Interval Information Criterion for the Two- and Three-parameter Logistic IRT Model

The formulas for the interval information criterion can be simplified for the two- and three-parameter logistic models. For the three-parameter logistic IRT-model the probability of a correct answer is

$$P_i(\theta) = c_i + (1-c_i)\frac{e^{a_i(\theta-b_i)}}{1+e^{a_i(\theta-b_i)}}, \tag{23}$$

and the corresponding item information function is given by

$$I_i(\theta) = \frac{a_i^2(1-c_i)}{\left(c_i+e^{a_i(\theta-b_i)}\right)\left(1+e^{-a_i(\theta-b_i)}\right)^2}, \tag{24}$$

where $a_i \in \mathbb{R}^+$, $b_i \in \mathbb{R}$ and $c_i \in \mathbb{R}^+$ are the discrimination, difficulty and guessing parameter, respectively and $\theta \in \mathbb{R}$ is the ability parameter. $\mathbb{R}$ and $\mathbb{R}^+$ are sets of real and positive real numbers, respectively.

For the three-parameter logistic model, the Interval Information Criterion in equation (14) can be reduced to (see Lord & Novick , 1968, p. 464 for an

outline of the proof):

$$\int_{\theta=\hat{\theta}_L}^{\hat{\theta}_R} I_i(\theta)d\theta = \frac{a_i}{1-c_i}\left[c_i \ln\left(\frac{P_i(\hat{\theta}_L)}{P_i(\hat{\theta}_R)}\right) + P_i(\hat{\theta}_R) - P_i(\hat{\theta}_L)\right]. \quad (25)$$

For the two-parameter logistic model ($c_i = 0$) the interval criterion becomes:

$$\int_{\theta=\hat{\theta}_L}^{\hat{\theta}_R} I_i(\theta)d\theta = a_i\,[P_i(\hat{\theta}_R) - P_i(\hat{\theta}_L)]. \quad (26)$$

When $\hat{\theta}_L = -\infty$ and $\hat{\theta}_R = \infty$ are used as starting values as suggested in equation (16), equation (25) for the three-parameter model will reduce to:

$$\int_{\theta=-\infty}^{\infty} I_i(\theta)d\theta = a_i\left[\frac{c_i \ln(c_i)}{1-c_i} + 1\right]. \quad (27)$$

and the same starting values for the two-parameter model will, of course, lead to:

$$\int_{\theta=-\infty}^{\infty} I_i(\theta)d\theta = a_i. \quad (28)$$

Equations (27) and (28) show that, when the three-parameter model holds instead of the two-parameter model, the mean amount of information is reduced by a factor $-c_i \ln(c_i) / (1-c_i)$. This reduction is considerable, even for relatively small values of the guessing parameter. For example, for $c_i = 0.2$ the reduction will amount to about 40%.

References

Apostol, T.M. (1967). Calculus volume 1, second edition. New York: John Wiley & Sons.

Berger, M.P.F. (to appear). A general approach to algorithmic design of fixed-form tests, adaptive tests and testlets. Applied Psychological Measurement. Accepted with revision.

Bock, R.D., & Mislevy, R.J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. Applied Psychological Measurement, 6, 431-444.

Bradley, R.A., & Gart, J.J. (1962). The asymptotic properties of ML estimators when sampling from associated populations. Biometrika, 49, 205-214.

Kingsbury, G.G., & Zara, A.R. (1989). Procedures for Selecting Items for Computerized Adaptive Tests. Applied Measurement in Education, 2, 359-375

Lord, F.M., & Novick, M.R. (1968). Statistical theories of mental test scores. Reading, Mass: Addison-Wesley.

Lord, F.M. (1971). Robbins-Monro procedures for tailored testing. Educational and Psychological Measurement, 31, 3-31.

Lord, F.M. (1977). A broad-range tailored test of verbal ability. Applied Psychological Measurement, 1, 95-100.

Lord, F.M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Lawrence Erlbaum.

Mood, A.M., Graybill, F.A., & Boes, D.C. (1974). Introduction to the theory of statistics. Singapore: McGraw-Hill.

Samejima, F. (1973). A comment on Birnbaum's three-parameter logistic model in the latent trait theory. Psychometrika, 38, 221-233.

Wainer, H. (1990). Computerized adaptive testing: a primer. Hillsdale, NJ: Lawrence Erlbaum.

Weiss, D.J. (1976) Adaptive testing research in Minnesota: Overview, recent results, and future directions. In C.L. Clark (Ed.), Proceedings of the First Conference on Computerized Adaptive Testing (pp.24-35). Washington, DC: United States Civil Service Commission.

Weiss, D.J. (1978). Proceedings of the 1977 Computerized Adaptive Testing Conference. Minneapolis: University of Minnesota.

Yen, W.M., Burket, G.R., & Sykes, R.C. (1991). Nonunique solutions to the likelihood equation for the three-parameter logistic model. Psychometrika, 56, 39-54.

Authors' Note

The authors wish to acknowledge the assistance of Wim M.M. Tielen with the simulation study. They also want to thank Wim J. van der Linden for reviewing an earlier version of this paper.

Figure captions

Figure 1. Weight functions of three item selection criteria.

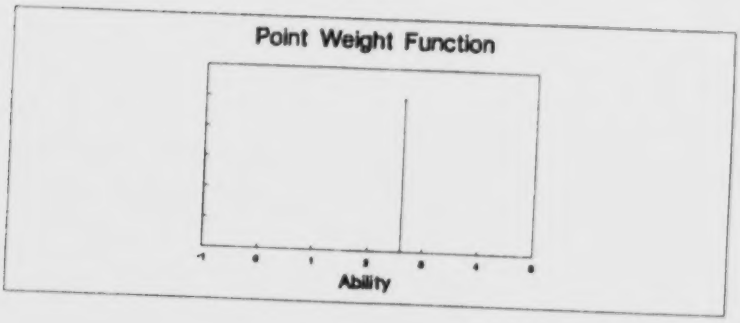
Figure 2. Mean average item information over 200 replications for different item selection criteria and ability values for the 200-item bank.

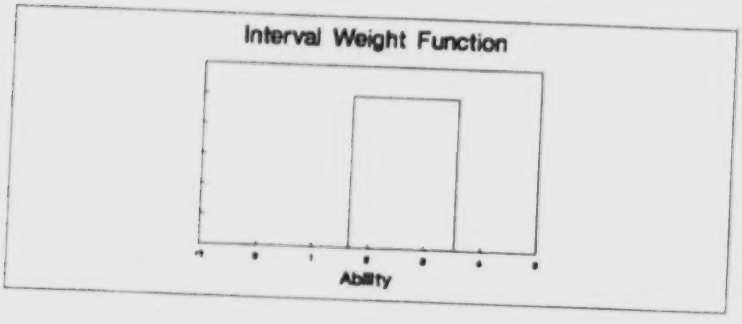
Figure 3. Mean average item information over 200 replications for different item selection criteria and ability values for the 400-item bank.

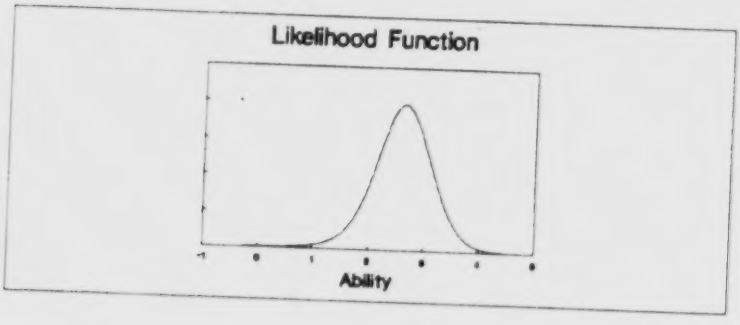
Figure 4. Minimum test information over 200 replications for different item selection criteria and ability values for the 200-item bank.

Figure 5. Minimum test information over 200 replications for different item selection criteria and ability values for the 400-item bank.

Figure 1

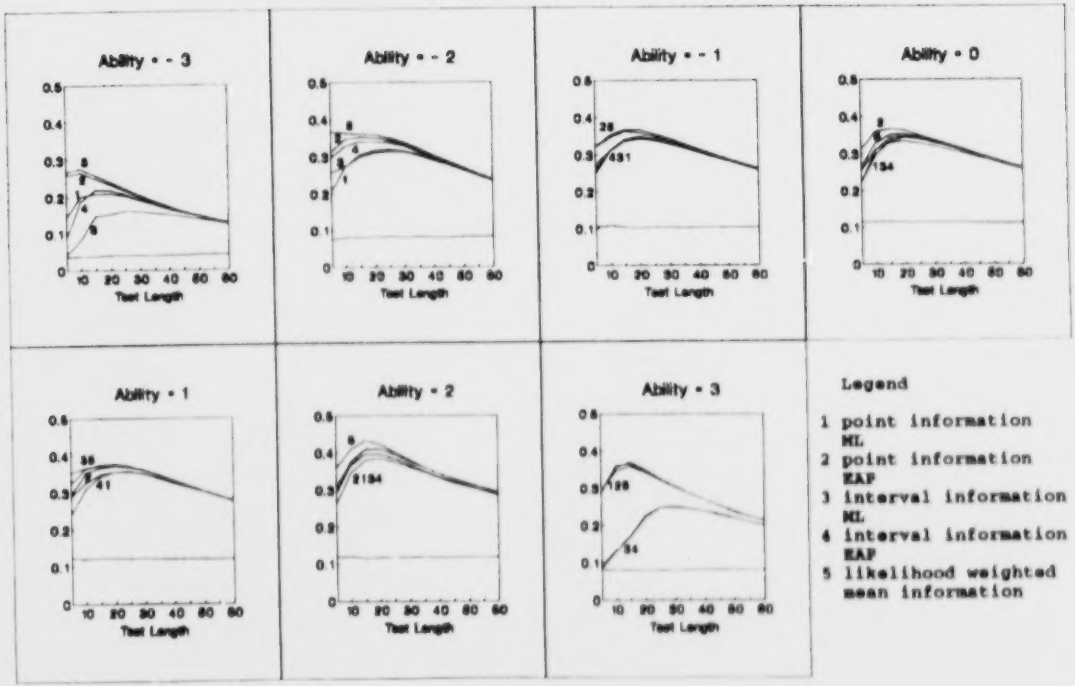






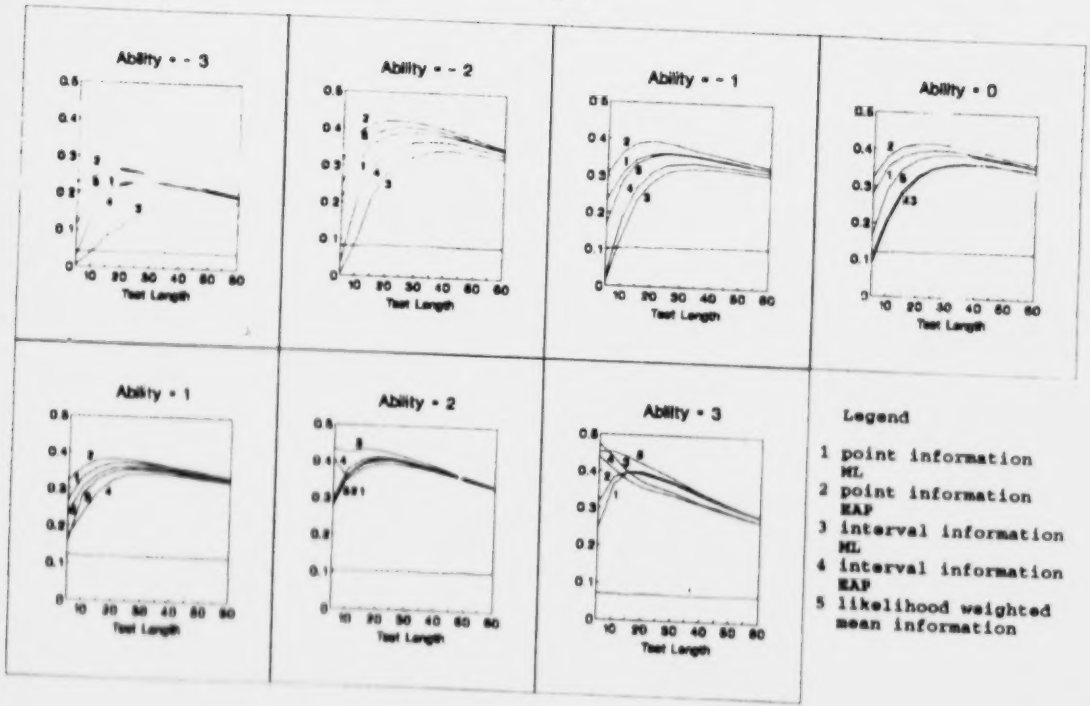
BEST COPY AVAILABLE

Figure 2



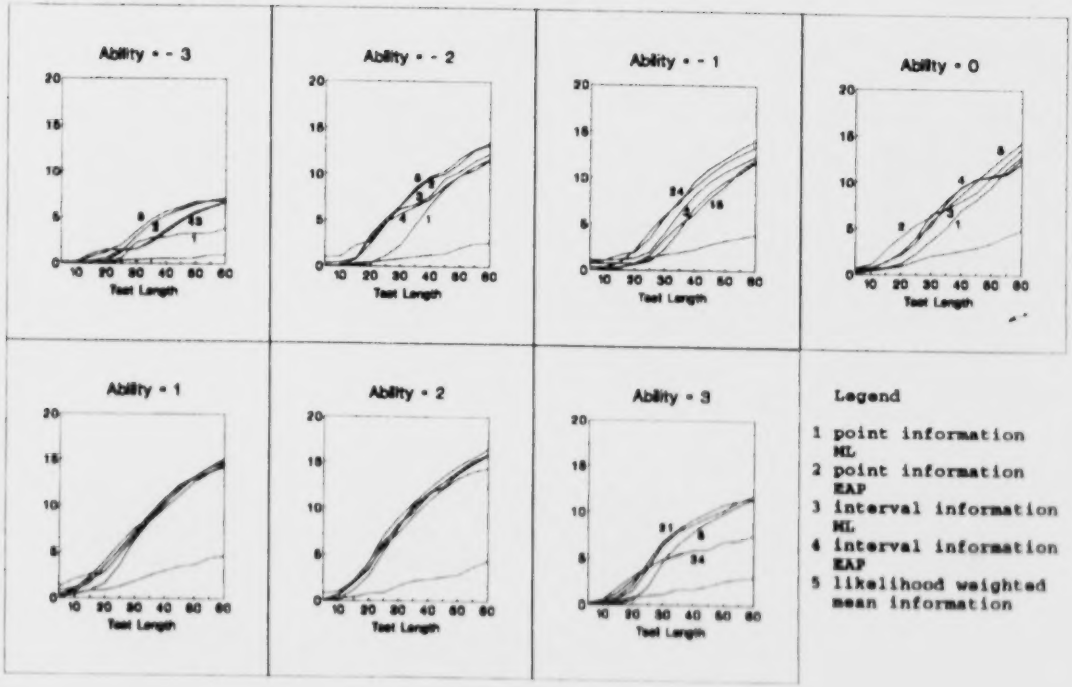
BEST COPY AVAILABLE

Figure 3



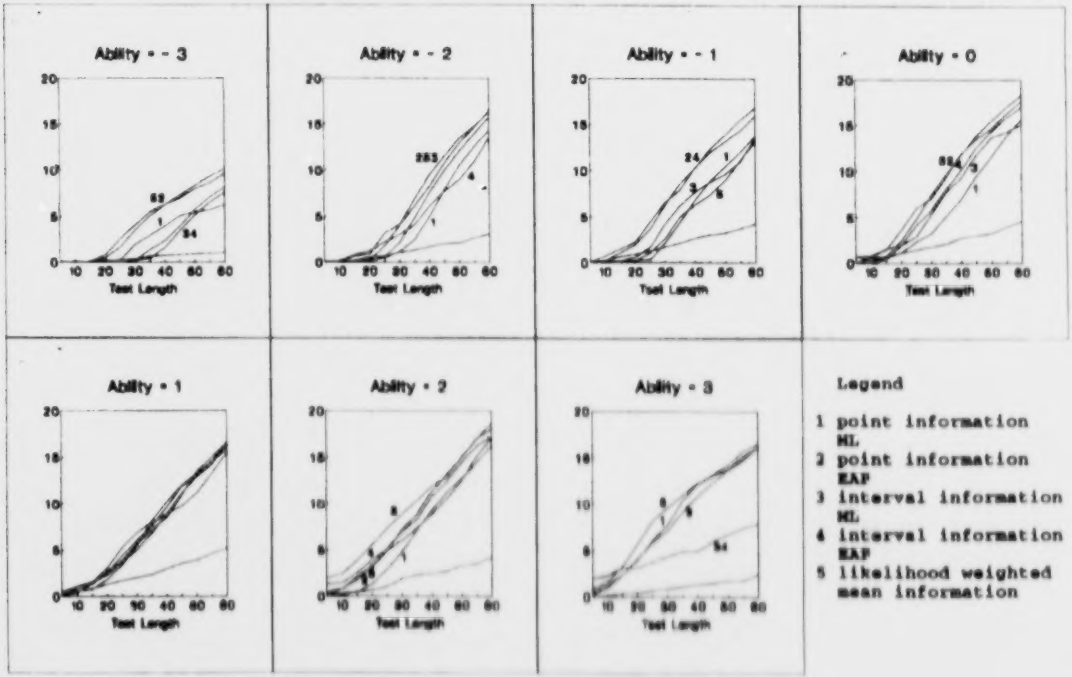
BEST COPY AVAILABLE

Figure 4



BEST COPY AVAILABLE

Figure 5



BEST COPY AVAILABLE

Titles of recent Research Reports from the Department of Educational Measurement and Data Analysis. University of Twente, Enschede, The Netherlands.

RR-94-6 W.J.J. Veerkamp & M.P.F. Berger, *Some new item selection criteria for adaptive testing*

RR-94-5 R.R. Meijer, K. Sijtsma & I.W. Molenaar, *Reliability estimation for single dichotomous items*

RR-94-4 M.P.F. Berger & W.J.J. Veerkamp, *A review of selection methods for optimal design*

RR-94-3 W.J. van der Linden, *A conceptual analysis of standard setting in large-scale assessments*

RR-94-2 W.J. van der Linden & H.J. Vos, *A compensatory approach to optimal selection with mastery scores*

RR-94-1 R.R. Meijer, *The influence of the presence of deviant item score patterns on the power of a person-fit statistic*

RR-93-1 P. Westers & H. Kelderman, *Generalizations of the Solution-Error Response-Error Model*

RR-91-1 H. Kelderman, *Computing Maximum Likelihood Estimates of Loglinear Models from Marginal Sums with Special Attention to Loglinear Item Response Theory*

RR-90-8 M.P.F. Berger & D.L. Knol, *On the Assessment of Dimensionality in Multidimensional Item Response Theory Models*

RR-90-7 E. Boekkooi-Timminga, *A Method for Designing IRT-based Item Banks*

RR-90-6 J.J. Adema, *The Construction of Weakly Parallel Tests by Mathematical Programming*

RR-90-5 J.J. Adema, *A Revised Simplex Method for Test Construction Problems*

RR-90-4 J.J. Adema, *Methods and Models for the Construction of Weakly Parallel Tests*

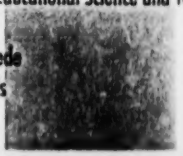
RR-90-2 H. Tobi, *Item Response Theory at subject- and group-level*

RR-90-1 P. Westers & H. Kelderman, *Differential item functioning in multiple choice items*

Research Reports can be obtained at costs from Bibliotheek, Faculty of Educational Science and Technology, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands.

faculty of
EDUCATIONAL SCIENCE
AND TECHNOLOGY

A publication by
The Faculty of Educational Science and Technology of the University of Twente
P.O. Box 217
7500 AE Enschede
The Netherlands



BEST COPY AVAILABLE